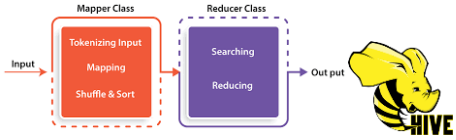
YOUTUVE DATA ANALYSIS

Big data and Hadoop

Problem Statement

A. Find out the top 5 categories with maximum number of videos uploaded.
B. Find out the top 10 rated videos.
C. Find out the most viewed videos.

Big Data Eco-system



Dataset description

Column1: Video id of 11 characters.
Column2: uploader of the video of string data type.
Column3: Interval between day of establishment of Youtube and the date of uploading of the video of integer data type.
Column4: Category of the video of String data type.
Column5: Length of the video of integer data type.
Column6: Number of views for the video of integer data type.
Column7: Rating on the video of float data type.
Column8: Number of ratings given on the video.
Column9: Number of comments on the videos in integer data type.
Column10: Related video ids with the uploaded video.

Table of Contents

Kaushik Dey
Kaushikdey1984@yahoo.com

Youtube data Analysis using Hive.

A. Login into Big Data Lab

Hi kaushik

We have set up labs for you so that you can practice all the assignments, case studies and projects mentioned under course content in your LMS.

Your BigData Lab is ready for use!

Please go through lab manual to know how to use the lab

Go to BigData Lab

B. Now, after logged in we can go to hive interface from that Bigdata lab, the screenshot is given below.

```
127 login: kaushikdey67edu
kaushikdey67edu@127.0.0.1's password:
Last login: Sun Mar 26 07:28:44 2023 from localhost
[kaushikdey67edu@ip-10-1-1-204 ~]$ hive
WARNING: Use "yarn jar" to launch YARN applications.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/Stat
icLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2023-03-26 10:40:15,025 main WARN JNDI lookup class is not available because this JRE does not support JNDI. JNDI string lookups will not be
 available, continuing configuration. Ignoring java.lang.ClassNotFoundException: org.apache.logging.log4j.core.lookup.JndiLookup

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/hive-common-2.1.1-cdh6.2.1.ja
r!/hive-log4j2.properties Async: false

WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive>
```

C. Now, we must write the HiveQL commands for that problem statement with step-by-step instructions.

1. **First, we must create database in hive environment.**

   Script:

   CREATE database youtubeAnalysis;

   Use youtubeAnalysis.

```
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create database youtubeDataAnalysis;
OK
Time taken: 1.618 seconds
hive> use youtubeDataAnalysis;
OK
Time taken: 0.178 seconds
hive>
```

2. **Second, we must create table inside the database with following scripts and necessary queries.**
   Script:

   ✓ CREATE TABLE if not exists youtubedata (vid string, upldr string, interval1 int, category string, length int, numview int, rating float,
   numratings int, numcomment int, relvids string)
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY '\t'
   STORED AS textfile.
   ✓ Show Tables.
   ✓ Describe youtubedata.

   youtubeanalysis.youtubedata

vid	string
upldr	string
interval1	int
category	string
length	int
numview	int
rating	float
numratings	int
numcomment	int
relvids	string

D. Now we must load data from NameNode to Hdfs. (in our case the file name is : youtubedata.txt.txt) via Hadoop commands or user interface. But Before that we have to upload our data to name node via FTP.

   ✓ **File upload via FTP to NameNode**

   ✓ **Load Data in hdfs via UI**
   Script:
   Hadoop fs -mkdir youtube_data
   Hadoop fs -put youtubedata.txt youtube_data
   Hadoop fs -ls youtube_data

Home / user / kaushikdey67edu / **youtube_data** Trash

Name	Size	User	Group	Permissions	Date
↑		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 26, 2023 05:15 AM
.		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 26, 2023 05:15 AM
youtubedata.txt	946.7 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 26, 2023 05:15 AM

Show 45 of 1 items Page 1 of 1

- ✓ **Load Data in hive from hdfs**

  Script

  LOAD DATA INPATH '/user/kaushikdey67edu/youtube_data/youtubedata.txt' INTO TABLE youtubedata;

  SELECT * FROM youtubedata LIMIT 4;

Image Screenshot from Hive Editor

Query History Saved Queries Results (4)

	youtubedata.vid	youtubedata.upldr	youtubedata.interval1	youtubed
1	QuRYeRnAuXM	EvilSquirrelPictures	1135	Pets & An
2	3TYqkBJ9YRk	hggh22	1135	Comedy
3	rSJ8QZWBegU	TimeGem	1135	Entertainr
4	nRcovJn9xHg	wooochacha	1135	Entertainr

Image Screenshot from Hive Shell

SELECT * FROM youtubedata;

bigdata.cloudlabs.edureka.co/shell/

TechnicalAuditRohit... CCA_28_Audit Daily_Status_Report Daily Attendance_s... Joining_Training_Co... Final Submission_st... Presentation Time...

```
LuyS9M8T03A     joeyfanatic     558     Music   254     446890  4.86    921     529     -7yyT7K-5jg
HqbVbPvlDoM     musicaldogsport 569     Pets & Animals  187     4049478 4.86    2955    945     NutzSiRpYBc
T2M6yV6mueg     enemigopublico  764     Music   316     1550140 4.82    3077    2609    tTaOvzZKRxA
MZm037jPNgc     NationalGeographic      909     Pets & Animals  258     699256  4.71    540     868     7ojhYBCnkuc
Wl_h586Ep9M     timsalmons1982  851     Music   306     328350  4.89    877     527     N9i2fqxSjTI
Lln5i1N3J8g     uncnick157      609     Music   240     474760  4.93    1743    1267    bN1ejG5R7mI
JnsOORuRB-M     Jasafar 461     Pets & Animals  379     45076   4.81    130     35      BRwsAXMseYI
We8P_Ww27hY     elpres42        789     Music   137     224770  4.9     639     378     _1Qo1eaWF8c
DGQVX8iGbgk     solasura        594     Pets & Animals  658     174587  4.73    1262    997     o-HEEW7YVNM
TnxQDrdIV2Q     raykay2 618     Entertainment   192     610733  4.73    1191    686     V2wmrbsDg9A
E4VMntSUskg     bchelms 584     Pets & Animals  512     327242  4.87    442     144     hYczb_I7QKk
GJ66UO-iKpc     rbsza   468     Music   343     309577  4.79    564     381     KrSs7gfLDjc
crmD_B8ERzk     freestyler343   293     Pets & Animals  257     289219  4.76    561     186     NutzSiRpYBc
LzI9VyIjpDI     machinima       1088    Entertainment   37      186036  4.82    310     677     e4yx1s-l6f8
UTYMb79BPGc     RockstarGames   1024    Entertainment   123     229527  4.83    816     1408    e4yx1s-l6f8
e8bEzU2Tx7c     machinima       1088    Entertainment   35      53490   4.82    117     254     LzI9VyIjpDI
vOskZNJJII8     GTASite 772     Entertainment   63      2708558 4.76    3810    5612    TecLbWk8A3w
_X5nHLOiAVo     machinima       1090    Entertainment   36      248579  4.66    524     683     LzI9VyIjpDI
EOj-iWkcJPU     Lassepc 863     Entertainment   63      370075  4.84    658     1246    HlF6fbIFiCM
M80K51DosFo     RockstarGames   772     Entertainment   63      437516  4.71    1083    1408    yB62b8S3Ra4
hKSk1bW9qzw     R2oor   863     Entertainment   63      745466  4.83    981     1955    vOskZNJJII8
5-reBn0gUkg     Ortmon  1064    Comedy  547     336956  4.58    2123    1384    e4yx1s-l6f8
47EWHY3E5AM     MgsTheFury404   1024    Entertainment   123     111371  4.77    215     422     e4yx1s-l6f8
buHrlMJu2Ss     RockstarGames   1016    Entertainment   48      108029  4.8     254     402     e4yx1s-l6f8
S95602a7cFE     mattCTB 1045    Entertainment   87      39885   4.51    63      131     UdYgoXeExWI
esCvu46qDaA     RSMFD   863     Entertainment   63      116929  4.84    182     222     8t4m5JM5r6Y
kOZ8bRAO7YQ     RockstarGames   NULL    UNA     63      199361  4.79    1256    2410    NULL
GSsVZcKHues     GTASite 1024    Entertainment   123     486752  4.86    1098    1764    TecLbWk8A3w
foKRqJGcbPQ     scrambledeggsTV 1100    Entertainment   136     13115   4.45    33      37      LzI9VyIjpDI
-CYhSNHbeC8     rehab07 771     Entertainment   186     103556  4.3     277     220     wtVeF5bT6YQ
Time taken: 0.25 seconds, Fetched: 4100 row(s)
```

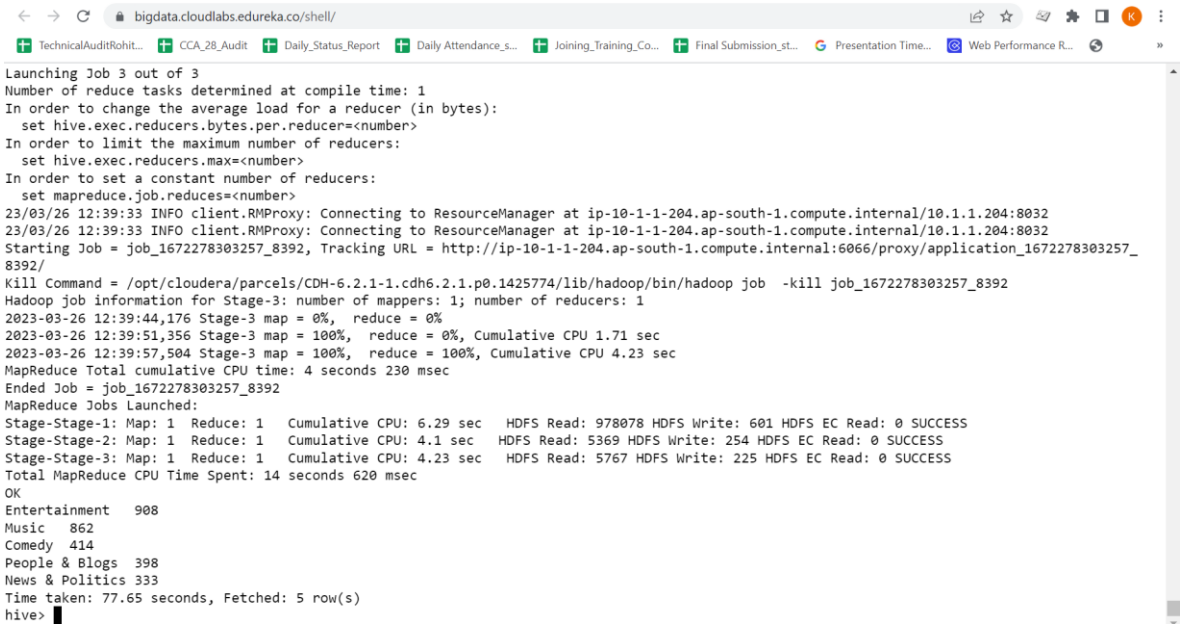
E. Problem Statement:

- ✓ Find out the top 5 categories with maximum number of videos uploaded.

Script:

Select category, count (*) as x from youtubedata group by category sort by x desc limit 5;

Image Screenshot from Hive Shell



```
Launching Job 3 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/03/26 12:39:33 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/03/26 12:39:33 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1672278303257_8392, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1672278303257_
8392/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1672278303257_8392
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2023-03-26 12:39:44,176 Stage-3 map = 0%,  reduce = 0%
2023-03-26 12:39:51,356 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.71 sec
2023-03-26 12:39:57,504 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 4.23 sec
MapReduce Total cumulative CPU time: 4 seconds 230 msec
Ended Job = job_1672278303257_8392
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.29 sec   HDFS Read: 978078 HDFS Write: 601 HDFS EC Read: 0 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.1 sec   HDFS Read: 5369 HDFS Write: 254 HDFS EC Read: 0 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 4.23 sec   HDFS Read: 5767 HDFS Write: 225 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 620 msec
OK
Entertainment   908
Music   862
Comedy  414
People & Blogs  398
News & Politics 333
Time taken: 77.65 seconds, Fetched: 5 row(s)
hive>
```

Category	Count
Entertainment	908
Music	862
Comedy	414
People & Blogs	398
News & Politics	333

Job Tracking Screenshots

Query | Search saved documents... | Jobs | kaushikdey67edu

Completed

Name	User	Type	Status	Progress	Group	Started	Duration	Id
Select category, count (*) as x from you...5 (Stage-3)	kaushikdey67edu	MAPREDUCE	SUCCEEDED	100%	root.default	March 26, 2023 6:09 PM	22.89s	application_1672278303257_8392
Select category, count (*) as x from you...5 (Stage-2)	kaushikdey67edu	MAPREDUCE	SUCCEEDED	100%	root.default	March 26, 2023 6:09 PM	23.87s	application_1672278303257_8391
Select category, count (*) as x from you...5	kaushikdey67edu	MAPREDUCE	SUCCEEDED	100%	root.default	March 26, 2023 6:08 PM	22.77s	application_1672278303257_8390

✓ Find out the top 10 rated videos.

Script:

Select vid, rating from youtubedata sort by rating desc limit 10;

Image Screenshot from Hive Shell

bigdata.cloudlabs.edureka.co/shell/

```
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/03/26 12:55:05 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/03/26 12:55:05 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1672278303257_8397, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1672278303257_
8397/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1672278303257_8397
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-03-26 12:55:14,612 Stage-2 map = 0%,  reduce = 0%
2023-03-26 12:55:22,865 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.38 sec
2023-03-26 12:55:29,062 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.53 sec
MapReduce Total cumulative CPU time: 5 seconds 530 msec
Ended Job = job_1672278303257_8397
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.45 sec   HDFS Read: 977082 HDFS Write: 426 HDFS EC Read: 0 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.53 sec   HDFS Read: 5920 HDFS Write: 367 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 980 msec
OK
3TYqkBJ9YRk     5.0
lD3GuCoJxT4     5.0
Y_hAXbgxXp8     5.0
gP0jnBrVEpI     5.0
YZev1imoxX8     5.0
L3mR8syHNIg     5.0
dh6dF1XY3uI     5.0
wzUyV42Izz4     5.0
osqnkdc349s     5.0
geUY_esOrt0     5.0
Time taken: 53.027 seconds, Fetched: 10 row(s)
hive>
```

VID	Ratings

3TYqkBJ9YRk	5.0
lD3GuCoJxT4	5.0
gP0jnBrVEpI	5.0
YZev1imoxX8	5.0
L3mR8syHNIg	5.0
dh6dF1XY3uI	5.0
wzUyV42Izz4	5.0
osqnkdc349s	5.0
geUY_esOrt0	5.0

- ✓ Find out the most viewed videos.

  Scripts:

  Select vid, numview from youtubedata sort by numview desc limit 100;

bigdata.cloudlabs.edureka.co/shell/

TechnicalAuditRohit... CCA_28_Audit Daily_Status_Report Daily Attendance_s... Joining_Training_Co... Final Submission_st... Presentation Time...

```
OK
12Z3J1uzd0Q     65341925
4DC4Rb9quKk     33754615
LU8DDYz68kM     27721690
kHmvkRoEowc     18235463
Md6rURKhZmA     18141492
EwTZ2xpQwpA     16841569
A2f3cuUXXRs     13038204
rZBA0SKmQy8     11007201
irp8CNj9qBI     10172172
ZCYaw5tGYAs     8944331
YCj-RyKCmHQ     8892789
Pxs5EM3xlus     8361412
eBGIQ7ZuuiU     8119804
rDB9zwlXrB8     7990144
Qit3ALTelOo     7939352
O61Do03ZCjw     7553998
veYbj2zypaM     7155928
iikKzQwgBJc     7043921
pFlcqWQVVuU     6618744
GfPJeDssBOM     6101232
Z_gKOCb4QBA     6024441
-hXKRilPNpc     5994026
EBM854BTGL0     5766247
HPPj6viIBmU     5565385
Ugrlzm7fySE     5360384
leEsz9ci5XE     5322084
yH8b5ruc_-E     5239520
Oy-nqyfR_w4     5205518
58CJih1iYC0     4915513
Qehxjub5lyo     4710964
cxbFLYa0_bw     4635992
```

bigdata.cloudlabs.edureka.co/shell/

TechnicalAuditRohit... CCA_28_Audit Daily_Status_Report Daily Attendance_s... Joining_Training_Co... Final Submission_st... Presentation Time... Web Pe

```
1GKaV2NDbuI     2464486
1EEynvjfljU     2446839
phL0RLKL8bc     2311640
WDxr-eakI6U     2298020
Wffwg7pA0t8     2192802
ichSb-1HSiY     2162891
IkmpIXd9Q90     2137442
clhOnZRrmXc     2135635
JhzysbVHX-o     2104830
0GMzzp_nJbw     2102333
tOUSGaUrHxc     2070346
V0t7MbeKVH8     2044382
3VVuMIB2hC0     2026565
rGrw4gWQmH8     1998552
5hayE_ZP0rI     1976909
Xv6lHwWwO3w     1965786
Dn_1ZWo0GNw     1897133
EnLI5UcNu5s     1867612
aC-KOYQsIvU     1838605
boc7rnhkLAk     1797772
udwUnwHJYck     1768978
8TNG5kAn1Dg     1762369
BPb0po2jzfg     1760933
_XC2mqcMMGQ     1760328
HlF6fbIFiCM     1743671
qsGQ10kN_U8     1725886
sNSxUqe91tA     1717986
4ira1eUOQGY     1714836
PP8avO4rlMQ     1700983
Ke1wTZQIAU4     1695836
9NxZy1GAF-I     1689873
NDbNb34fHJ0     1682470
le1i447wzPk     1670956
```

Youtube data Analysis using Map-reduce Job

Problem Statement A:

Find out the top 5 categories with maximum number of videos uploaded.

A. **Map-reduce job for top 5 Categories:**
First, we must create Top5_categories.java file under maven project.
1. **Code Top5_categories.java:**

```
package com.youtube;
//import the necessary libraries
import java.io.IOException;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
 //Top5_categories class
 public class Top5_categories {
    //Mapper class
    public static class MyMapper extends Mapper<LongWritable, Text, Text,
IntWritable> {
        private Text category = new Text();
        private final static IntWritable one = new IntWritable(1);
        public void map(LongWritable key, Text value, Context context )
        throws IOException, InterruptedException {
            String line = value.toString();
            String str[]=line.split("\t");
            if(str.length > 5){
                category.set(str[3]);
            }
        context.write(category, one);
    }
 }
 //Reducer class
    public static class MyReducer extends Reducer<Text, IntWritable,Text,
IntWritable> {
        public void reduce(Text key, Iterable<IntWritable> values, Context
context)
        throws IOException, InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
```

```
        context.write(key, new IntWritable(sum));
 }
 }
 // this is my main class
 public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(conf, "youtube_analysis_top_5_categories");
    job.setJarByClass(Top5_categories.class);
    job.setMapOutputKeyClass(Text.class);
    job.setMapOutputValueClass(IntWritable.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    // mapper & reducer class
    job.setMapperClass(MyMapper.class);
    job.setReducerClass(MyReducer.class);
    // input format & output format
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    // input & output location
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    //exit from that jar after completion
    System.exit(job.waitForCompletion(true) ? 0:1);
    }
 }
```

2. **Configuration for pom.xml:**

```
<?xml version="1.0" encoding="UTF-8"?>

<project xmlns="http://maven.apache.org/POM/4.0.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0
http://maven.apache.org/xsd/maven-4.0.0.xsd">
  <modelVersion>4.0.0</modelVersion>

  <groupId>com.youtube</groupId>
  <artifactId>data_analysis</artifactId>
  <version>1.0</version>


  <name>data_analysis</name>
  <!-- FIXME change it to the project's website -->
  <url>http://www.example.com</url>

  <properties>
    <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
    <maven.compiler.source>1.7</maven.compiler.source>
    <maven.compiler.target>1.7</maven.compiler.target>
```

```
  </properties>

  <dependencies>
    <dependency>
      <groupId>junit</groupId>
      <artifactId>junit</artifactId>
      <version>4.11</version>
      <scope>test</scope>
    </dependency>
    <dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-core</artifactId>
    <version>1.2.1</version>
</dependency>
<dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-common</artifactId>
    <version>2.7.2</version>
</dependency>
  </dependencies>

  <build>
    <pluginManagement><!-- lock down plugins versions to avoid using Maven
defaults (may be moved to parent pom) -->
      <plugins>
        <!-- clean lifecycle, see https://maven.apache.org/ref/current/maven-
core/lifecycles.html#clean_Lifecycle -->
        <plugin>
          <artifactId>maven-clean-plugin</artifactId>
          <version>3.1.0</version>
        </plugin>
        <!-- default lifecycle, jar packaging: see
https://maven.apache.org/ref/current/maven-core/default-
bindings.html#Plugin_bindings_for_jar_packaging -->
        <plugin>
          <artifactId>maven-resources-plugin</artifactId>
          <version>3.0.2</version>
        </plugin>
        <plugin>
          <artifactId>maven-compiler-plugin</artifactId>
          <version>3.8.0</version>
        </plugin>
        <plugin>
          <artifactId>maven-surefire-plugin</artifactId>
          <version>2.22.1</version>
        </plugin>
        <plugin>
          <artifactId>maven-jar-plugin</artifactId>
```

```
          <version>3.0.2</version>
        </plugin>
        <plugin>
          <artifactId>maven-install-plugin</artifactId>
          <version>2.5.2</version>
        </plugin>
        <plugin>
          <artifactId>maven-deploy-plugin</artifactId>
          <version>2.8.2</version>
        </plugin>
        <!-- site lifecycle, see https://maven.apache.org/ref/current/maven-core/lifecycles.html#site_Lifecycle -->
        <plugin>
          <artifactId>maven-site-plugin</artifactId>
          <version>3.7.1</version>
        </plugin>
        <plugin>
          <artifactId>maven-project-info-reports-plugin</artifactId>
          <version>3.0.0</version>
        </plugin>
      </plugins>
    </pluginManagement>
  </build>
</project>
```

3. **Now we must create the jar file. the location of jar file.**

This PC > Local Disk (D:) > Big_data > Big_Data_Hadoop_Certification_Training_Course > Youtube_Data_Analysis > youtbb > jar

Name	Date modified	Type	Size
top_5.jar	28-03-2023 08:21	JAR File	32,003 KB

4. **Now we must submit our hadoop jar file through web Shell.**
   A. Screen shot for Hadoop file location from Hue Portal.

Home / user / kaushikdey67edu / **youtube_data** Trash

Name	Size	User	Group	Permissions	Date
↑		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 26, 2023 05:38 AM
.		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 27, 2023 07:56 PM
top_cat		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 27, 2023 07:56 PM
youtubedata.txt	946.7 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 26, 2023 05:15 AM

Show 45 of 2 items Page 1 of 1

B. The code for Hadoop map reduce creation with the following commands.

hadoop jar top_5.jar '/user/kaushikdey67edu/youtube_data/youtubedata.txt' '/user/kaushikdey67edu/youtube_data/top_cat'

```
[kaushikdey67edu@ip-10-1-1-204 ~]$ hadoop jar top_5.jar '/user/kaushikdey67edu/youtube_data/youtubedata.txt' '/user/kaushikdey67edu/youtube_
data/top_cat'
WARNING: Use "yarn jar" to launch YARN applications.
23/03/28 03:16:47 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/03/28 03:16:48 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and exe
cute your application with ToolRunner to remedy this.
23/03/28 03:16:48 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/kaushikdey67edu/.staging/job_1672278303257_84
56
23/03/28 03:16:48 INFO input.FileInputFormat: Total input files to process : 1
23/03/28 03:16:48 INFO mapreduce.JobSubmitter: number of splits:1
23/03/28 03:16:48 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.sys
tem-metrics-publisher.enabled
23/03/28 03:16:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1672278303257_8456
23/03/28 03:16:48 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/03/28 03:16:49 INFO conf.Configuration: resource-types.xml not found
23/03/28 03:16:49 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/03/28 03:16:49 INFO impl.YarnClientImpl: Submitted application application_1672278303257_8456
23/03/28 03:16:49 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1672
278303257_8456/
23/03/28 03:16:49 INFO mapreduce.Job: Running job: job_1672278303257_8456
23/03/28 03:16:58 INFO mapreduce.Job: Job job_1672278303257_8456 running in uber mode : false
23/03/28 03:16:58 INFO mapreduce.Job:  map 0% reduce 0%
23/03/28 03:17:05 INFO mapreduce.Job:  map 100% reduce 0%
23/03/28 03:17:18 INFO mapreduce.Job:  map 100% reduce 80%
23/03/28 03:17:19 INFO mapreduce.Job:  map 100% reduce 100%
23/03/28 03:17:19 INFO mapreduce.Job: Job job_1672278303257_8456 completed successfully
23/03/28 03:17:19 INFO mapreduce.Job: Counters: 54
```

Now we must check the map reduce job from hue portal.

Home / user / kaushikdey67edu / youtube_data / **top_cat** Trash

Name	Size	User	Group	Permissions	Date
↑		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 27, 2023 08:16 PM
.		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 27, 2023 08:17 PM
_SUCCESS	0 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 27, 2023 08:17 PM
part-r-00000	18 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 27, 2023 08:17 PM
part-r-00001	86 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 27, 2023 08:17 PM
part-r-00002	42 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 27, 2023 08:17 PM
part-r-00003	33 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 27, 2023 08:17 PM
part-r-00004	78 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 27, 2023 08:17 PM

C. The commands for top 5 categories with following commands.

hadoop fs -cat /user/kaushikdey67edu/youtube_data/top_cat/part-r-00004 | sort -n -k2 -t$'\t' -r | head -n5

```
[kaushikdey67edu@ip-10-1-1-204 ~]$ hadoop fs -cat /user/kaushikdey67edu/youtube_data/top_cat/part-r-00004 | sort -n -k2 -t$'\t' -r | head -n
5
Entertainment	911
Music	870
People & Blogs	399
News & Politics	343
Sports	253
[kaushikdey67edu@ip-10-1-1-204 ~]$
```

Problem Statement B:

Find out the top 10 rated videos.

1. **Map-reduce job for top 10 rated videos:**
   First, we must create Top5_categories.java file under maven project.
   Code Top10_rated_videos.java:

```
package com.youtube;
//import the necessary libraries
import java.io.IOException;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class Top10_rated_videos {
    // create the top 10 rated video mapper
    public static class Top10VideoMap extends Mapper<LongWritable, Text, Text,
FloatWritable> {
        private Text video_name = new Text();
        private FloatWritable rating = new FloatWritable();

        public void map(LongWritable key, Text value, Context context)
                throws IOException, InterruptedException {
            String line = value.toString();
            String str[] = line.split("\t");
            if (str.length > 7) {
                video_name.set(str[0]);
                if (str[6].matches("\\d+(\\.\\d{1,3})?")) {
                    float f = Float.parseFloat(str[6]); // type casting string
to float
                    rating.set(f);
                }
            }
            context.write(video_name, rating);
        }
    }

    // top 10 video reducer
    public static class Top10VideoReduce extends Reducer<Text, FloatWritable,
Text, FloatWritable> {
        public void reduce(Text key, Iterable<FloatWritable> values,
                Context context)
                throws IOException, InterruptedException {
            float sum = 0;
```

```
            int l = 0;
            for (FloatWritable val : values) {
                l += 1;
                sum += val.get();
            }
            sum = sum / l;
            context.write(key, new FloatWritable(sum));
        }
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Top_10_video_rating");
        job.setJarByClass(Top10_rated_videos.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(FloatWritable.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(FloatWritable.class);
        job.setMapperClass(Top10VideoMap.class);
        job.setReducerClass(Top10VideoReduce.class);
        // input format & output format
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        // input & output location
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        //exit from that jar after completion
        System.exit(job.waitForCompletion(true) ? 0:1);
        }

}
```

2. **Now we must create the jar file. the location of jar file.**

Name	Date modified	Type	Size
top_5_categories.jar	28-03-2023 08:21	JAR File	32,003 KB
top_10_rated_video.jar	28-03-2023 12:58	JAR File	32,006 KB

3. **Now we must submit our hadoop jar file through web Shell.**
   A. Screen shot for Hadoop file location from Hue Portal.

Home / user / kaushikdey67edu / **youtube_data** Trash

Name	Size	User	Group	Permissions	Date
↑		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 26, 2023 05:38 AM
.		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 12:33 AM
top_10_rated_video		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 12:33 AM
top_cat		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 27, 2023 08:17 PM
youtubedata.txt	946.7 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 26, 2023 05:15 AM

Show 45 of 3 items Page 1 of 1

B. The code for Hadoop map reduce creation with the following commands.

hadoop jar top_10_rated_video.jar '/user/kaushikdey67edu/youtube_data/youtubedata.txt' '/user/kaushikdey67edu/youtube_data/top_10_rated_video'

```
[kaushikdey67edu@ip-10-1-1-204 ~]$ hadoop jar top_10_rated_video.jar '/user/kaushikdey67edu/youtube_data/youtubedata.txt' '/user/kaushikdey6
7edu/youtube_data/top_10_rated_video'
WARNING: Use "yarn jar" to launch YARN applications.
23/03/28 07:33:09 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/03/28 07:33:10 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and exe
cute your application with ToolRunner to remedy this.
23/03/28 07:33:10 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/kaushikdey67edu/.staging/job_1672278303257_84
61
23/03/28 07:33:10 INFO input.FileInputFormat: Total input files to process : 1
23/03/28 07:33:10 INFO mapreduce.JobSubmitter: number of splits:1
23/03/28 07:33:10 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.sys
tem-metrics-publisher.enabled
23/03/28 07:33:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1672278303257_8461
23/03/28 07:33:10 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/03/28 07:33:10 INFO conf.Configuration: resource-types.xml not found
23/03/28 07:33:10 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/03/28 07:33:10 INFO impl.YarnClientImpl: Submitted application application_1672278303257_8461
23/03/28 07:33:10 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1672
278303257_8461/
23/03/28 07:33:10 INFO mapreduce.Job: Running job: job_1672278303257_8461
23/03/28 07:33:19 INFO mapreduce.Job: Job job_1672278303257_8461 running in uber mode : false
23/03/28 07:33:19 INFO mapreduce.Job:  map 0% reduce 0%
23/03/28 07:33:26 INFO mapreduce.Job:  map 100% reduce 0%
23/03/28 07:33:45 INFO mapreduce.Job:  map 100% reduce 20%
23/03/28 07:33:47 INFO mapreduce.Job:  map 100% reduce 60%
23/03/28 07:33:48 INFO mapreduce.Job:  map 100% reduce 100%
23/03/28 07:33:48 INFO mapreduce.Job: Job job_1672278303257_8461 completed successfully
23/03/28 07:33:48 INFO mapreduce.Job: Counters: 54
```

Now we must check the map reduce job from hue portal.

Home / user / kaushikdey67edu / youtube_data / **top_10_rated_video** Trash

Name	Size	User	Group	Permissions	Date
↑		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 12:33 AM
.		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 12:33 AM
_SUCCESS	0 bytes	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 28, 2023 12:33 AM
part-r-00000	12.9 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 28, 2023 12:33 AM
part-r-00001	13.5 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 28, 2023 12:33 AM
part-r-00002	13.6 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 28, 2023 12:33 AM
part-r-00003	13.2 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 28, 2023 12:33 AM
part-r-00004	12.4 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 28, 2023 12:33 AM

The commands for top 10 rated videos.

hadoop fs -cat /user/kaushikdey67edu/youtube_data/top_10_rated_video/part-r-00004 | sort -n -k2 -t$'\t' -r | head -n10

```
127 login: kaushikdey67edu
kaushikdey67edu@127.0.0.1's password:
Last login: Tue Mar 28 07:32:40 2023 from localhost
[kaushikdey67edu@ip-10-1-1-204 ~]$ hadoop fs -cat /user/kaushikdey67edu/youtube_data/top_10_rated_video/part-r-00004 | sort -n -k2 -t$'\t' -
r | head -n10
ZRMYeOVn8ew     5.0
ZozTCX4LLn8     5.0
zokUPp_nj_c     5.0
zNxXy-tO7l4     5.0
ZLZIv0FYwF8     5.0
ZLG-SrZoxI4     5.0
ZgcUkiqCxpo     5.0
ZboNI4DGsmg     5.0
Za0bJiONMd4     5.0
Z-2Ye_7lpt4     5.0
[kaushikdey67edu@ip-10-1-1-204 ~]$
```

Problem Statement C:

Find out the most viewed videos.

1. **Map-reduce job for most viewed Videos:**

First, we must create Most_viewed_videos.java file under maven project.

- ✓ **Code Most_viewed_videos.java:**

```
package com.youtube;

//import the necessary libraries
import java.io.IOException;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class Most_viewed_videos {
    // Most Viewed Mapper
    public static class MostViewedMap extends Mapper<LongWritable, Text, Text,
IntWritable> {
        private Text vidid = new Text();
        private IntWritable numview = new IntWritable();

        public void map(LongWritable key, Text value, Context context)
                throws IOException, InterruptedException {
            String line = value.toString();
            String str[] = line.split("\t");
            if (str.length > 5) {
                vidid.set(str[0]);
                if (str[5].matches("\\d+.+")) {
                    int f = (int) Float.parseFloat(str[5]);
                    numview.set(f);
```

```
                }
            }
            context.write(vidid, numview);
        }
    }
    // Most Viewed Reducer
    public static class MostViewedReduce extends Reducer<Text, IntWritable,
Text, IntWritable> {
        public void reduce(Text key, Iterable<IntWritable> values,
                Context context) throws IOException, InterruptedException {
            int sum = 0;
            int l = 0;
            for (IntWritable val : values) {
                l += 1; // counts number of values are there for that key
                sum += val.get();
            }
            sum = sum / l; // takes the average of the sum
            context.write(key, new IntWritable(sum));
        }
    }
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Most_Viewed_video");
        job.setJarByClass(Most_viewed_videos.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(IntWritable.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        //set Mapper Name
        job.setMapperClass(MostViewedMap.class);
        //set Reducer Name
        job.setReducerClass(MostViewedReduce.class);
        // input format & output format
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        // input & output location
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        //exit from that jar after completion
        System.exit(job.waitForCompletion(true) ? 0:1);
        }

}
```

- ✓ **Now we must create the jar file. the location of jar file.**

most_viewed_videos.jar	28-03-2023 13:35	JAR File	32,010 KB
top_5_categories.jar	28-03-2023 08:21	JAR File	32,003 KB
top_10_rated_video.jar	28-03-2023 12:58	JAR File	32,006 KB

- ✓ **Now we must submit our hadoop jar file through web Shell.**

Home / user / kaushikdey67edu / **youtube_data** Trash

Name	Size	User	Group	Permissions	Date
↑		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 26, 2023 05:38 AM
.		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 01:07 AM
most_viewed_videos		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 01:08 AM
top_10_rated_video		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 28, 2023 12:33 AM
top_cat		kaushikdey67edu	kaushikdey67edu	drwxr-xr-x	March 27, 2023 08:17 PM
youtubedata.txt	946.7 KB	kaushikdey67edu	kaushikdey67edu	-rw-r--r--	March 26, 2023 05:15 AM

Show 45 of 4 items Page 1 of 1

- ✓ **The code for Hadoop map reduce creation with the following commands.**

hadoop jar most_viewed_videos.jar '/user/kaushikdey67edu/youtube_data/youtubedata.txt' '/user/kaushikdey67edu/youtube_data/most_viewed_videos'

```
[kaushikdey67edu@ip-10-1-1-204 ~]$ hadoop jar most_viewed_videos.jar '/user/kaushikdey67edu/youtube_data/youtubedata.txt' '/user/kaushikdey6
7edu/youtube_data/most_viewed_videos'
WARNING: Use "yarn jar" to launch YARN applications.
23/03/28 08:07:45 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/03/28 08:07:45 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and exe
cute your application with ToolRunner to remedy this.
23/03/28 08:07:45 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/kaushikdey67edu/.staging/job_1672278303257_84
62
23/03/28 08:07:46 INFO input.FileInputFormat: Total input files to process : 1
23/03/28 08:07:46 INFO mapreduce.JobSubmitter: number of splits:1
23/03/28 08:07:46 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.sys
tem-metrics-publisher.enabled
23/03/28 08:07:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1672278303257_8462
23/03/28 08:07:46 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/03/28 08:07:46 INFO conf.Configuration: resource-types.xml not found
23/03/28 08:07:46 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/03/28 08:07:46 INFO impl.YarnClientImpl: Submitted application application_1672278303257_8462
23/03/28 08:07:46 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1672
278303257_8462/
23/03/28 08:07:46 INFO mapreduce.Job: Running job: job_1672278303257_8462
23/03/28 08:07:56 INFO mapreduce.Job: Job job_1672278303257_8462 running in uber mode : false
23/03/28 08:07:56 INFO mapreduce.Job:  map 0% reduce 0%
23/03/28 08:08:02 INFO mapreduce.Job:  map 100% reduce 0%
23/03/28 08:08:18 INFO mapreduce.Job:  map 100% reduce 20%
23/03/28 08:08:19 INFO mapreduce.Job:  map 100% reduce 100%
23/03/28 08:08:19 INFO mapreduce.Job: Job job_1672278303257_8462 completed successfully
```

The commands for most viewed Videos.

hadoop fs -cat /user/kaushikdey67edu/youtube_data/most_viewed_videos/part-r-00004 | sort -n -k2 -t$'\t' -r | head -n50

```
[kaushikdey67edu@ip-10-1-1-204 ~]$ hadoop fs -cat /user/kaushikdey67edu/youtube_data/most_viewed_videos/part-r-00004 | sort -n -k2 -t$'\t' -
r | head -n50
ZCYaw5tGYAs     8944331
Qit3ALTelOo     7939352
Z_gKOCb4QBA     6024441
-hXKRilPNpc     5994026
Ugrlzm7fySE     5360384
ylPqlrBoyRY     4388073
m2OC5Z1Fii8     4372739
AIiMa2Fe-ZQ     3695341
2JC3UMJ2It4     3470648
_pGvEGFkvNs     3209797
hNxhrPaaCA4     3062136
8hvEsCUj0xQ     2790875
vOskZNJJII8     2708558
3t8GdtYdRk0     2657650
1EEynvjfljU     2446839
Wffwg7pA0t8     2192802
0GMzzp_nJbw     2102333
5hayE_ZP0rI     1976909
Xv6lHwWwO3w     1965786
aC-KOYQsIvU     1838605
BPb0po2jzfg     1760933
HlF6fbIFiCM     1743671
qsGQ10kN_U8     1725886
PP8avO4rlMQ     1700983
9NxZy1GAF-I     1689873
le1i447wzPk     1670956
552k0ifjuDw     1598694
```

The 3 jobs screenshots are given below.

youtube_analysis_top_5_categories	kaushikdey67edu	MAPREDUCE	SUCCEEDED	100%	root.default	March 28, 2023 8:46 AM	29.43s	application_167
Top_10_video_rating	kaushikdey67edu	MAPREDUCE	SUCCEEDED	100%	root.default	March 28, 2023 1:03 PM	36.3s	application_167
Most_Viewed_video	kaushikdey67edu	MAPREDUCE	SUCCEEDED	100%	root.default	March 28, 2023 1:37 PM	31.26s	application_167

Thank You